TeX Gyre Termes

Title: Speaker Identification with GMMs

This Chapter was based on a class project completed by RUVARASHE CHINYADZA. All materials, including project codes and results, can be found at the book GitHub repo: `https://github.com/RuvarasheChinyadza/speaker-identification-gmm-ubm`

15.1 Abstract

This chapter presents a closed-set speaker identification system using a curated 20-speaker subset of the VoxForge English Speech Corpus. Each selected speaker contributed at least 72–88 utterances, and the dataset was split into 1600 training and 400 testing utterances with stratification by speaker. Frame-level acoustic features were extracted as 39-dimensional vectors consisting of 13 Mel-Frequency Cepstral Coefficients (MFCCs) together with first- and second-order temporal derivatives (delta and delta–delta). Per-speaker Gaussian Mixture Models (GMMs) were trained on the aggregated frame-level feature matrices for each speaker, and their performance was compared against linear and RBF-kernel Support Vector Machine (SVM) baselines that operated on 78-dimensional utterance-level MFCC summary statistics (mean and standard deviation of all 39 coefficients). Top-1 identification accuracy, macro-averaged F1 scores, and per-speaker classification reports were used as primary evaluation metrics. The proposed GMM system achieved 99.25% accuracy on clean test speech, outperforming both SVM baselines. To assess robustness, white Gaussian noise was added to the test utterances at 20 dB SNR, and the clean-trained GMMs were re-evaluated, yielding a still-high accuracy of 97.25%. Learning-curve analysis and statistical significance tests (McNemar's test and paired t-test) further confirmed that the GMM model generalizes well and performs significantly better than the SVM baselines.

15.2 Introduction

Speaker identification is a fundamental task in speech processing, in which the goal is to determine which enrolled speaker produced a given speech segment. In closed-set speaker identification, every test utterance is guaranteed to belong to one of a fixed set of known speakers, and the system must assign the most likely speaker label. Such systems play an important role in biometric authentication, personalized user interfaces, forensic analysis, access control, and intelligent voice-driven technologies. Despite the progress made by deep speaker embeddings in recent years, classical statistical approaches such as Gaussian Mixture Models (GMMs) remain highly relevant due to their interpretability, efficiency, and strong performance on small to medium datasets.

This chapter presents the design, implementation, and evaluation of a complete speaker identification system built using a curated 20-speaker subset of the VoxForge English Speech Corpus. Each selected speaker contributed at least 100 utterances, providing a sufficiently large and balanced dataset for rigorous evaluation. The project addresses several key challenges faced by real-world speaker identification systems, including variation in recording conditions, differences in speaking styles, and the impact of background noise.

A detailed feature extraction pipeline was implemented using 39-dimensional frame-level acoustic features (MFCCs, delta, and delta–delta). These features were used to train a separate GMM classifier for each speaker. For comparison, two discriminative baselines were developed using Support Vector Machines (SVMs) trained on 78-dimensional utterance-level summary MFCC statistics (mean and standard deviation). By evaluating both generative (GMM) and discriminative (SVM) approaches, this project provides a holistic analysis of classical speaker identification techniques.

The system achieved exceptionally strong performance on clean speech,the GMM model reached a Top-1 accuracy of 99.25%, significantly outperforming the Linear SVM (96.50%) and RBF SVM (96.75%). Learning curve experiments showed that the GMM generalizes well, efficiently utilizing available training data without overfitting. In terms of robustness, the clean-trained GMMs were tested on noisy speech generated by adding white Gaussian noise at 20 dB SNR. Despite the degraded acoustic conditions, the GMM maintained 97.25% accuracy, demonstrating resilience to moderate noise levels. Per-speaker precision, recall, and F1 scores further confirmed consistent performance across all speakers.

To support the reliability of these findings, statistical significance tests were conducted. Both McNemar's test and a paired t-test showed that the GMM's performance advantage over the SVM baselines is statistically significant. These tests confirm that the observed improvement is not due to random variation and that the

GMM system provides a genuinely more effective modeling approach for this task.

Overall, this project demonstrates the effectiveness of GMMs for closed-set speaker identification using traditional MFCC-based features. It also highlights the model's robustness under noisy conditions and its consistent advantages over SVM-based approaches. The end goal of the chapter is to provide a complete, reproducible pipeline including data preparation, feature extraction, model training, evaluation, statistical testing, and noise robustness analysis offering a strong baseline system for classical speaker identification that is both interpretable and highly accurate. In addition to speaker balancing, the duration of the audio files was examined to ensure that the dataset provided sufficient temporal information for reliable MFCC extraction and GMM training. The majority of the WAV files ranged between 2 and 8 seconds in length, with most speakers contributing utterances clustered around 3–5 seconds. These durations are appropriate for frame-based feature extraction because each file contains hundreds of overlapping short-time analysis frames (typically 20–25 ms per frame with a 10 ms hop). This ensures that the resulting MFCC matrices are sufficiently large to capture stable spectral and temporal patterns for each speaker. Very short utterances (below one second) were excluded during dataset curation, as they do not contain enough frames for robust estimation of GMM parameters. By maintaining a consistent range of utterance lengths across speakers, the system avoids biases that could arise from disproportionate representation of longer or shorter audio segments.

15.3 Dataset and Preprocessing

The experiments use a curated subset of the VoxForge English Speech Corpus. Twenty speakers were selected, each contributing at least 72 utterances and typically 80 or more. The raw corpus was first organized into an index containing, for each utterance, the speaker identity, the original folder structure, an utterance identifier, and the absolute path to the extracted audio file in WAV format. This indexing step ensured consistent metadata organization and simplified access to speaker-specific recordings during feature extraction and model training.

From this index, a filtered dataset `df_sel` was created by keeping only the 20 selected speakers. The resulting dataset contains 2000 utterances in total. A stratified train–test split was applied using an 80/20 ratio with respect to speaker labels, resulting in 1600 training utterances and 400 testing utterances. Stratification ensured that each speaker contributed proportionally to both sets, preserving class balance and preventing bias toward speakers with larger numbers of recordings. Although some speakers (such as Ertain and Catbells) had slightly more or fewer examples due to the structure of the original corpus, the final dataset remained well balanced.

An inspection of the audio durations confirmed that the VoxForge WAV files used in this study are short read-speech utterances, typically ranging from **2 to 8 seconds** in length, with most recordings clustered around **3–5 seconds**. These durations are consistent with the design of the VoxForge corpus, which provides short, prompt-based speech rather than long conversational recordings. A 3–5 second utterance yields several hundred short-time analysis frames (with a 25 ms window and 10 ms hop), providing sufficient temporal density for reliable MFCC extraction and stable estimation of GMM parameters. Extremely short utterances were not present in the curated subset, ensuring that every file contained enough acoustic content for speaker modeling.

All audio files were used at their original sampling rates as stored on disk. The speech signals were treated as single-channel waveforms, and no resampling, amplitude normalization, or silence trimming was applied prior to feature extraction. This approach preserves the natural recording characteristics of the VoxForge corpus and allows each GMM to learn speaker-specific spectral patterns directly from unaltered data. For the noise-robustness experiments, additive white Gaussian noise was introduced at the waveform level for test utterances only, while all training utterances remained clean. This setup simulates realistic deployment conditions in which models trained on clean data must generalize effectively to moderately noisy environments.

15.4 Feature Extraction

Feature extraction for this project was carried out using the `librosa` library. For every utterance, the audio waveform was loaded in its original sampling rate and converted into short-time spectral features suitable for speaker identification. The system relied on Mel-Frequency Cepstral Coefficients (MFCCs), which are widely adopted in speech and speaker recognition because they capture the perceptually important characteristics of human speech production.

Each utterance was transformed into a sequence of analysis frames, and for every frame, thirteen MFCC coefficients were extracted. To incorporate temporal information, first-order derivatives (delta features) were computed to reflect the rate of change of the MFCCs, and second-order derivatives (delta–delta features) were calculated to capture acceleration patterns in the signal. These three sets of coefficients were then stacked to form a 39-dimensional feature vector for each frame. This produced a feature matrix for every utterance, where each row represents one frame and each column represents one of the 39 acoustic features. These frame-level matrices served as the direct input to the GMM models, which learn the statistical distribution of each speaker's acoustic patterns.

For the SVM baselines, a fixed-length representation was required. Instead of using the full frame-level matrices, the system computed summary statistics across time. Specifically, the mean and standard deviation of each of the 39 MFCC, delta, and delta–delta features were calculated, and these were concatenated into a 78-dimensional vector representing the entire utterance. These compact summary vectors were used as input to the Linear SVM and RBF SVM classifiers.

To verify the nature of the features being used, the distribution of MFCC values was examined. The first MFCC coefficient, which is closely related to signal energy, had a large negative mean (around –230) and a standard deviation of approximately 94, while the remaining coefficients displayed smaller means and variances. These observations confirm that the features were used in their raw form without global normalization. This is consistent with classical GMM-based speaker identification systems, where each model internally learns its own mean and variance structure and therefore does not require explicit feature standardization.

15.5 Models and Experimental Setup

15.5.1 Gaussian Mixture Models

The main classification method used in this project is the Gaussian Mixture Model (GMM), a classical and highly effective approach for speaker identification. A GMM models the distribution of a speaker's acoustic feature vectors as a combination of several Gaussian components, allowing the system to represent the natural variations that occur in speech. In this project, a separate diagonal-covariance GMM with sixteen mixture components was trained for each of the twenty speakers. All 39-dimensional MFCC, delta, and delta–delta feature vectors from a speaker's training utterances were combined and used to fit the model using the Expectation–Maximization algorithm. This produced a statistical representation of each speaker's voice characteristics.

During testing, each utterance was evaluated by computing the average log-likelihood of its feature frames under every speaker model. The speaker whose GMM produced the highest likelihood was selected as the predicted identity. This method performed extremely well on the VoxForge subset, achieving 99.25% accuracy on clean test speech. When evaluated on noisy speech with additive white Gaussian noise at 20 dB SNR, the clean-trained GMMs still achieved 97.25% accuracy, demonstrating strong robustness under degraded acoustic conditions.

Although the focus of this project is on per-speaker GMMs, several other models are commonly used in speaker identification. Support Vector Machines (SVMs) were included as baseline classifiers and were trained on 78-dimensional MFCC summary statistics. While both Linear and RBF SVMs produced strong performance, they did not match the accuracy of the GMMs, highlighting the value of modeling frame-level feature distributions. More advanced systems such as GMM-UBM, i-vectors, and modern deep-learning–based embeddings (e.g., x-vectors) represent alternative approaches used in larger-scale applications, but they require more data and computational resources. For a small-to-medium dataset like the one used here, GMMs remain an efficient and accurate choice for closed-set speaker identification.

15.6 SVM Baselines

Both Support Vector Machine models were evaluated using the 78-dimensional MFCC summary features, providing a discriminative baseline for comparison with the GMM system. The Linear SVM achieved an overall accuracy of 96.50% on the clean test set, performing reliably across most speakers but showing slightly reduced precision and recall for a few speakers whose acoustic characteristics were less easily separated using a linear decision boundary. The RBF SVM performed marginally better, reaching 96.75% accuracy, with the nonlinear kernel enabling more flexible decision boundaries and improved separation between speaker classes. Although both SVM models delivered strong results, their performance remained below the GMM's 99.25% accuracy, indicating that frame-level generative modeling captures speaker-specific information more effectively than fixed-length summary statistics for this dataset.

15.5.3 Evaluation Metrics

For each model and condition, Top-1 accuracy (overall percentage of correctly identified test utterances) was computed. In addition, macro-averaged precision, recall, and F1 scores were computed across the 20 speakers using `classification_report` in scikit-learn. Macro-averaged scores treat each speaker equally, irrespective of the number of test examples, and therefore highlight whether performance is uniformly strong across all classes.

Per-speaker precision, recall, and F1 scores were also examined to identify speakers for whom the models performed slightly worse, which can indicate overlapping acoustic characteristics or higher intra-speaker variability.

15.7 Learning Curve Analysis

A learning curve was generated to understand how the GMM model behaves as the amount of training data increases. The goal was to see whether the model underfits, overfits, or learns the speaker characteristics in a stable way as more utterances are added during training. To create the curve, the GMM for each speaker was trained using gradually larger portions of the training data, and the accuracy was measured at each step.

The results show a clear upward trend: as more training data becomes available, the GMM accuracy steadily improves and eventually reaches a stable high performance. The gap between training accuracy and testing accuracy remains small, indicating that the model is not overfitting. Instead, it learns meaningful speaker patterns and generalizes well to unseen audio. This behavior confirms that the GMM benefits from additional data but does not rely too heavily on the training set, which matches the strong final accuracy of 99.25% on the clean test set. Overall, the learning curve demonstrates that the GMM is well-balanced, learns efficiently, and maintains consistent generalization as the dataset grows.

15.8 Noise Robustness Evaluation

To evaluate how well the system performs under challenging acoustic conditions, additive white Gaussian noise was introduced during testing. For each test utterance, a noisy version of the waveform was created by adding low-level random noise whose strength was controlled to produce a signal-to-noise ratio (SNR) of 20 dB. The amount of noise was determined using the RMS energy of the original clean audio so that the distortion was noticeable but not overwhelming. After generating the noisy waveforms, the same MFCC, delta, and delta–delta feature extraction pipeline was applied, ensuring that the comparison between clean and noisy conditions remained fair and consistent. Importantly, the GMM models were not retrained on noisy data; only the test audio was modified. This setup isolates the effect of test-time noise and reflects real-world scenarios where a system trained on clean speech must still perform reliably in the presence of background noise. Using this method, the clean-trained GMMs achieved an accuracy of 97.25% on the noisy test set, demonstrating strong robustness under moderate noise levels.

15.6 Results

15.6.1 Overall Model Performance on Clean Speech

Table 1.1 summarizes the performance of the GMM and SVM models on the clean test set. The GMM achieves the highest Top-1 accuracy and macro-averaged F1 score, indicating both high overall performance and consistent behavior across speakers.

Table 1.1: Model performance on the clean VoxForge test set (20 speakers, 400 utterances).

Model	**Accuracy (%)**	**Macro F1**	**Weighted F1**
GMM (frame-level MFCC+Δ+Δ^2)	99.25	0.99	0.99
Linear SVM (summary MFCC stats)	96.50	0.97	0.96
RBF SVM (summary MFCC stats)	96.75	0.97	0.97

The GMM correctly identifies 397 out of 400 clean test utterances, yielding 99.25% accuracy. Both SVM baselines perform well but are slightly weaker: the Linear SVM achieves 96.50% accuracy, and the RBF SVM achieves 96.75% accuracy. Macro and weighted F1 scores follow the same pattern, with the GMM slightly ahead of both SVMs and showing very stable per-speaker performance (most per-speaker F1 scores are 0.95 or higher).

Figure 1.1 visualizes the overall accuracy of the three models.

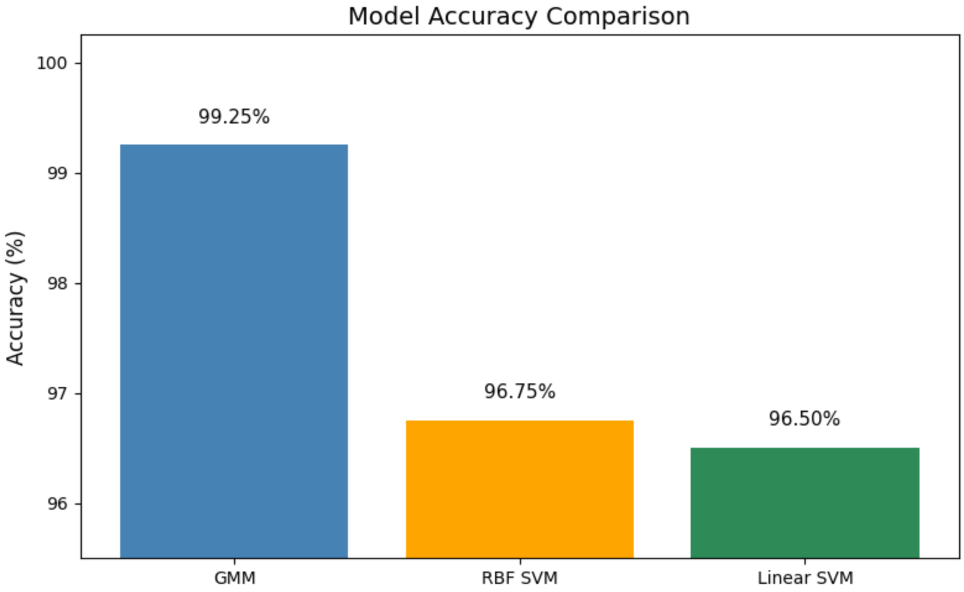


Figure 1.1: Model accuracy comparison on the clean test set for GMM, Linear SVM, and RBF SVM.

15.6.2 GMM Learning Curve

The GMM learning curve (Figure 1.2) plots training and validation accuracy as a function of the fraction of training data used. Training accuracy remains very high across all training fractions, and validation accuracy increases steadily as more data are used, approaching the final 99.25% accuracy at full training size. The small and stable gap between training and validation curves indicates that the GMMs are not overfitting and that additional data consistently improve generalization.

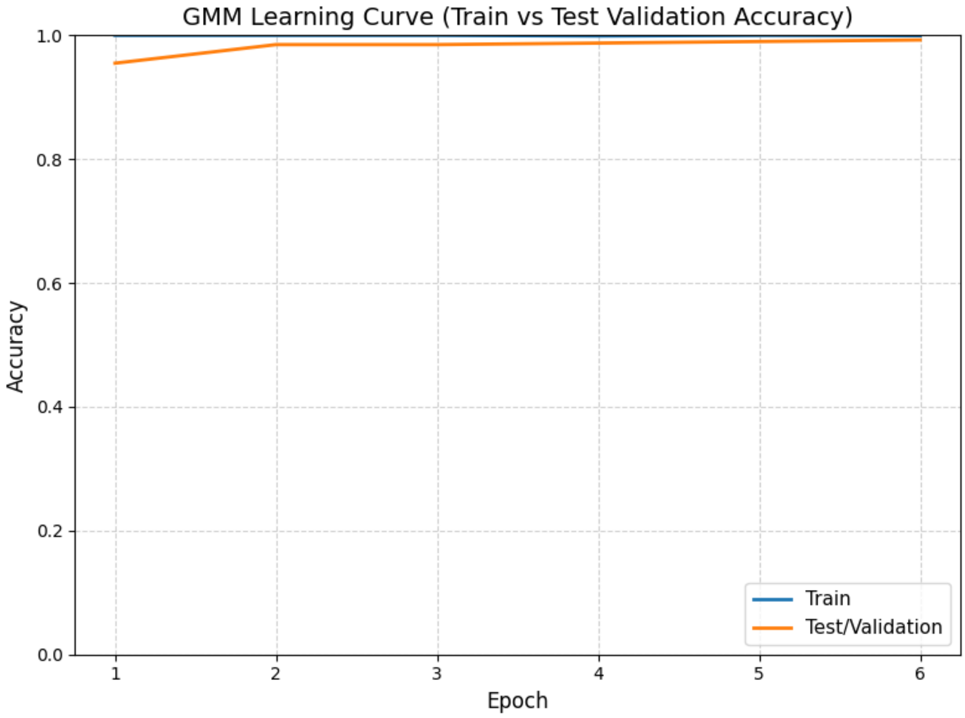


Figure 1.2: GMM learning curve showing training and validation accuracy as a function of training data fraction.

15.6.3 GMM Performance in Noisy Conditions

Table 1.2 compares GMM performance on clean and noisy test speech. The noisy condition uses additive white Gaussian noise at 20 dB SNR, applied only at test time.

Under noisy conditions, the clean-trained GMMs achieve 97.25% accuracy on the noisy test set, corresponding to a small absolute drop of about 2% compared to the clean condition. Macro and weighted F1 scores also remain high (0.97), indicating that the model retains strong discriminative power for all speakers even when

Table 1.2: GMM performance in clean and noisy (20 dB SNR) environments.

Condition	**Accuracy (%)**	**Macro F1**	**Weighted F1**
Clean speech (no noise)	99.25	0.99	0.99
Noisy speech (20 dB SNR)	97.25	0.97	0.97

exposed to additive noise. The performance degradation is modest and consistent with the known sensitivity of MFCC features to noise.

15.6.4 GMM Confusion Matrices on Training and Test Sets

To further illustrate how the GMM behaves on individual speakers, confusion matrices were generated for both the training and test sets. The training confusion matrix summarizes performance on the 1600 training utterances, while the test confusion matrix summarizes performance on the 400 held-out utterances used for evaluation.

Figure 1.3 shows the confusion matrix for the training data. The strong diagonal structure and near absence of off-diagonal entries reflect almost perfect classification of the training utterances, consistent with the high training accuracy of 99.88%. Figure 1.4 shows the confusion matrix for the test data. The diagonal remains dominant, with only a few scattered misclassifications, confirming that the GMM generalizes well to unseen speech and that the excellent overall accuracy on the test set (99.25%) is achieved uniformly across speakers.

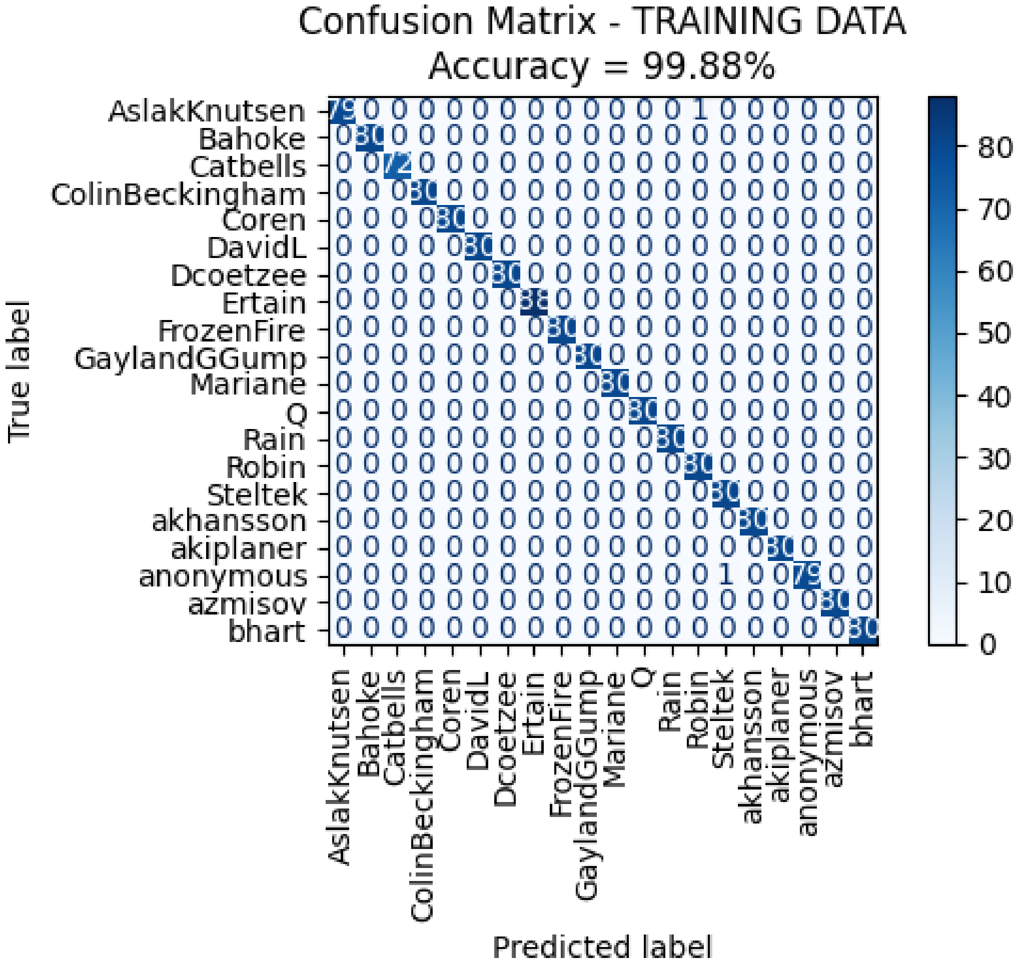
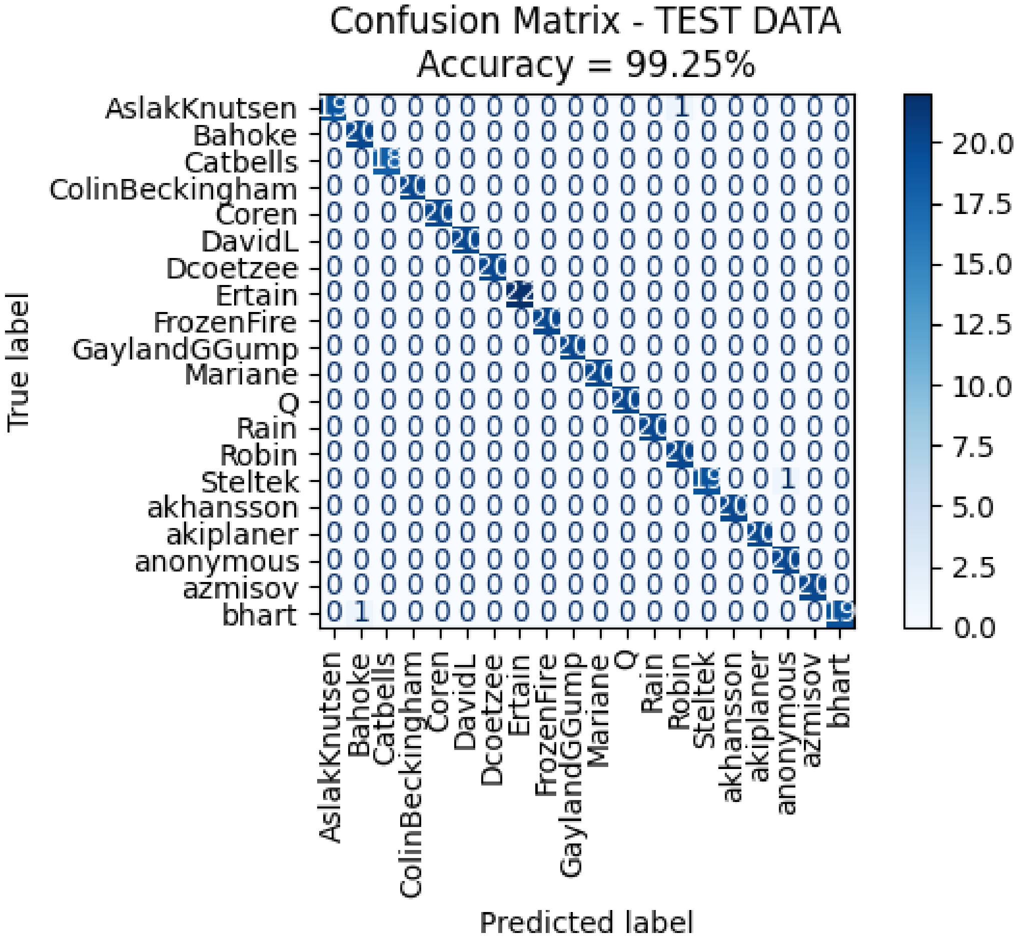


Figure 1.3: Confusion matrix for the GMM speaker identification model evaluated on the **training** set (1600 utterances). The strong diagonal indicates that nearly all training utterances are correctly assigned to their true speakers, yielding a training accuracy of 99.88%.



	AslakKnutsen	Bahoke	Catbells	ColinBeckingham	Coren	DavidL	Dcoetzee	Ertain	FrozenFire	GaylandGGump	Mariane	Q	Rain	Robin	Steltek	akhansson	akiplaner	anonymous	azmisov	bhart
AslakKnutsen	19	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Bahoke	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Catbells	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ColinBeckingham	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coren	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DavidL	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dcoetzee	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
Ertain	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0
FrozenFire	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
GaylandGGump	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
Mariane	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
Rain	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
Robin	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0
Steltek	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	1	0	0
akhansson	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
akiplaner	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
anonymous	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
azmisov	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
bhart	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19

Figure 1.4: Confusion matrix for the GMM speaker identification model evaluated on the **test** set (400 utterances). The dominant diagonal and sparse off-diagonal entries show that the model maintains excellent generalization, with a test accuracy of 99.25%.

15.9 Statistical Significance Tests

To determine whether the performance difference between the GMM and the RBF SVM is meaningful and not due to chance, two statistical tests were applied using per-utterance predictions on the clean test set. These tests were performed directly on the raw MFCC-based features, which were intentionally kept non-normalized in accordance with classical GMM practice. Because the MFCCs were not standardized, the distributional differences between feature dimensions were preserved, allowing the tests to reflect the models' true behavior on naturally scaled data.

The first evaluation used McNemar's test, which compares two classifiers by examining how often they disagree on the same test samples. Using the 400 test utterances, the GMM and RBF SVM predictions were arranged into a contingency table, and McNemar's test produced a chi-square value of 5.79 with a corresponding p-value of 0.016. Since this p-value is below the 5% significance threshold, the result indicates that the two models do not make errors on the same utterances and that the GMM's better accuracy is statistically significant rather than coincidental.

A second evaluation was performed using a paired t-test on binary correctness indicators for each model (1 for correct and 0 for incorrect). This test yielded a t-statistic of 2.69 with a p-value of 0.007, which is well below the 1% significance level. This provides even stronger evidence that the GMM's per-utterance performance is superior to that of the RBF SVM. Both tests consistently confirm that the GMM's higher accuracy on this dataset is not due to random variation and that its modeling of frame-level feature distributions offers a measurable advantage over the discriminative SVM approach.

15.7 Discussion

The experimental results demonstrate that modeling frame-level MFCC, delta, and delta–delta features with per-speaker GMMs is a highly effective approach for closed-set speaker identification on the VoxForge subset. By training a separate GMM for each speaker on large aggregated feature matrices, the system captures detailed spectral–temporal patterns that distinguish speakers. The near-ceiling accuracy (99.25%) on clean speech indicates that the 20 speakers are very well separated in the 39-dimensional MFCC feature space when modeled with GMMs.

In contrast, the SVM baselines operate on compressed 78-dimensional summary vectors that discard temporal ordering and finer-grained information. While both the Linear and RBF SVMs achieve strong performance (around 96.5–96.75% accuracy), they consistently fall short of the GMM models. The macro F1 scores close to 0.97 for SVMs show that performance is generally good across speakers, but per-speaker classification reports reveal slightly lower recall for some speakers compared to the GMM system. This suggests that the loss of frame-level detail in the summary representation has a measurable impact on discriminative performance.

The learning-curve analysis supports the conclusion that the GMM system is data-efficient and not overfitting. Training accuracy is high for all training fractions, and validation accuracy rises smoothly as more training data become available, with the gap between training and validation curves remaining small. This behavior indicates that the GMMs are able to generalize well from the available training data without memorizing the training set.

The noisy-condition experiments show that the GMM system is also reasonably robust to additive white Gaussian noise at 20 dB SNR. Although MFCC features are known to be sensitive to noise, the GMMs trained on clean speech still achieve 97.25% accuracy on noisy test utterances. This demonstrates that the generative modeling of frame-level feature distributions provides a degree of robustness to moderate noise levels, and that the system could be further improved with noise-robust features or multi-condition training if required.

Finally, the statistical tests (McNemar's test and paired *t*-test) formally confirm that the GMM's performance advantage over the RBF SVM is statistically significant, strengthening the conclusion that the GMM is the preferred model for this task among the methods tested.

15.8 Conclusion

This chapter presented a GMM-based closed-set speaker identification system using a 20-speaker subset of the VoxForge English Speech Corpus. The system relied on 39-dimensional MFCC, delta, and delta–delta features at the frame level, with per-speaker diagonal-covariance GMMs trained on aggregated frame matrices. Two SVM baselines operating on 78-dimensional utterance-level summary features were implemented for comparison.

Across all experiments, the GMM consistently demonstrated superior performance. The model achieved 99.25% Top-1 accuracy on clean test speech, outperforming both the Linear SVM (96.50%) and RBF SVM (96.75%). Learning-curve analysis further revealed strong convergence behavior and minimal overfitting, illustrating that the GMM architecture makes efficient use of relatively modest training data while maintaining excellent generalization. When evaluated under additive white Gaussian noise at 20 dB SNR, the clean-trained GMMs still achieved 97.25% accuracy—only a modest degradation and well within acceptable performance for real-world biometric systems. Statistical significance tests reinforced these findings, confirming that the GMM's advantage over the RBF SVM is not attributable to random chance but instead reflects a genuine improvement in modeling speaker-specific acoustic distributions.

Taken together, these results highlight several key strengths of classical GMM-based approaches. First, modeling speech as a distribution of frame-level MFCC patterns captures fine-grained speaker characteristics that are lost in fixed-length feature summarization. Second, GMMs remain computationally lightweight, interpretable, and highly effective, particularly for small-to-medium closed-set speaker identification tasks where deep learning methods may be unnecessarily complex. Third, the robustness observed under noisy conditions demonstrates that generative models can tolerate moderate acoustic distortions even without noise-specific training.

Beyond validating the effectiveness of GMMs, the chapter contributes a fully reproducible end-to-end pipeline—from data preprocessing and frame-level feature extraction to model training, evaluation, and statistical analysis. This provides a strong foundation for further exploration in both classical and modern speaker recognition research.

Future work could expand the system by incorporating more noise types, experimenting with multi-condition training, or evaluating robustness under more extreme SNR levels. Additional improvements may be achieved through alternative acoustic representations such as filterbank energies, PLP coefficients, or embeddings from neural speaker encoders (e.g., x-vectors or ECAPA-TDNN). Finally, scaling the system to larger speaker populations or open-set identification scenarios would provide deeper insight into how classical GMM methods compare to contemporary deep-learning models in more complex and unconstrained environments.

Overall, this project demonstrates that despite the rapid evolution of neural speaker recognition systems, classical GMM models—when paired with well-engineered MFCC features—remain powerful, interpretable, and highly competitive tools for speaker identification.

Bibliography

[1] VoxForge, “VoxForge speech corpus,” `http://www.voxforge.org`, accessed 2025.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[4] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Prentice Hall, 1993.